Informació addicional pel projecte de PROP
Quadrimestre de tardor, curs 25/26

Extractor de prototipus de comportament o perfils

Un possible joc de proves podrien ser les enquestes del CIS:

https://www.kaggle.com/code/amcamc/encuesta-bar-metro-cis-66010323

Representació de les dades

S'utilitzarà el model d'espai vectorial per a representar les dades. El conjunt de *N* respostes d'una persona *i* es representarà com un vector $X_i = [X_{i1}, X_{i2}, \dots, X_{iN}]$. Així doncs, el conjunt de totes les respostes de totes les *M* persones serà una matriu $X = \begin{bmatrix} X_{11} & \cdots & X_{1N} \\ \vdots & \ddots & \vdots \\ X_{M1} & \cdots & X_{MN} \end{bmatrix}$ on cada resposta a una pregunta ($X_{.j}$) no és més que una variable d'un cert tipus. Les variables que modelitzen les respostes podran ser:

- *Variables quantitatives o numèriques*, que emmagatzemen un sol valor numèric.
- *Variables qualitatives ordenades*, que emmagatzemen un sol valor o modalitat qualitativa (i.e. molt-poc/poc/normal/força/molt). Les diferents modalitats possibles són ordenades.
- *Variables qualitatives no ordenades, que emmagatzemen un sol valor o modalitat qualitativa* (i.e. groc/blau/verd/vermell/lila/marró). Les diferents modalitats possibles no són ordenades.
- *Variables qualitatives no ordenades, que emmagatzemen un conjunt de p valors o modalitats qualitatives d'un màxim de q modalitats diferents* ($1 \leq p \leq q$) (i.e. {groc, verd, lila}). Les diferents modalitats possibles no són ordenades.
- *Variables string en format lliure*, que emmagatzemen un sol valor que és un string qualsevol. No existeix cap valor predeterminat.

Mesures de semblança o distàncies

Els algorismes de clustering consisteixen fonamentalment en una sèrie de processos iteratius, que es basen en càlculs de semblances o distàncies entre els diferents objectes que es volen agrupar (respostes de les persones en el nostre cas), o entre centroides dels clústers o grups o entre ambdós tipus d'elements.
Els conceptes de semblança i distància són inversos i habitualment els seus valors estan normalitzats en l'interval [0,1]. Així la semblança entre dues persones X_i i X_j es pot calcular a partir de la seva distància i viceversa:

$$Semb(X_i, X_j) = 1 - D(X_i, X_j)$$

En el cas de que els objectes siguin vectors la distància entre dos objectes és la agregació de les distàncies locals entre cada component del mateix. És a dir, es pot expressar la distància entre dues respostes de persones X_i i X_j com:

$$D(X_i, X_j) = \sum_{k=1}^{N} \frac{w_k * D_{local}(X_{ik}, X_{jk})}{\sum_{k=1}^{N} w_k} \qquad on \;\; 0 \leq \; w_k \leq 1$$

on w_k representa el pes o importància (*weight*) d'aquella variable. De moment, podeu treballar pensant que totes les respostes (variables) tenen el mateix pes. Llavors, és trivial veure que el càlcul de la distància es transforma en:

$$D(X_i, X_j) = \sum_{k=1}^{N} \frac{D_{local}(X_{ik}, X_{jk})}{N}$$

La distància global sempre pertany a l'interval [0,1], ja que les distàncies locals també pertanyen a l'interval [0,1]. Aquesta fórmula per la D correspon a l'anomenada distància L1, o Manhattan normalitzada. Alternativament, es pot utilitzar també la distància Euclídea, que es calcula de la següent manera:

$$D(X_i, X_j) = \frac{\sqrt{\sum_{k=1}^{N} D_{local}(X_{ik}, X_{jk})^2}}{N}$$

Podeu escollir la funció de distància que considereu més adient pel vostre conjunt de dades. Depenent del cas, és possible que considereu implementar i utilitzar més d'una.

El càlcul de les distàncies locals de cada component del vector (de cada resposta) depèn del tipus de variable que emmagatzema la resposta. Per tant, això comporta que la distància global utilitzada sigui de tipus heterogènia. Les distàncies locals (normalitzades a l'interval [0,1]) més habituals que s'utilitzen i que us recomanem utilitzar són:

1. $D_{local}(X_{ik}, X_{jk}) = \frac{|X_{ik} - X_{jk}|}{V_{max}(X_{.k}) - V_{min}(X_{.k})}$

   $on\ \ V_{max}(X_{.k})\ \ i\ \ V_{min}(X_{.k})\ \ són\ els\ valors\ màxim\ i\ mínim\ de\ la\ variable\ X_{.k}$

   quan la variable $X_{.k}$ és numèrica o quantitativa

2. $D_{local}(X_{ik}, X_{jk}) = \frac{|mod(X_{ik}) - mod(X_{jk})|}{\#mod - 1}$

   $on\ \ mod(X_{ik})$ és el numeral corresponent a l'ordinal que li correspon a aquesta modalitat, després d'ordenar totes les modalitats existents,
   i $\#mod$ és el nombre de modalitats existents

   quan la variable $X_{.k}$ és qualitativa ordenada i emmagatzema un sol valor

3. $D_{local}(X_{ik}, X_{jk}) = \begin{cases} 0 & si\ X_{ik} = X_{jk} \\ 1 & si\ X_{ik} \neq X_{jk} \end{cases}$

   quan la variable $X_{.k}$ és qualitativa no ordenada i emmagatzema un sol valor

4. $D_{local}(X_{ik}, X_{jk}) = 1 - Jaccard(X_{ik}, X_{jk}) = 1 - \frac{\#(\{X_{ik}\} \cap \{X_{jk}\})}{\#(\{X_{ik}\} \cup \{X_{jk}\})}$

   on $\{X_{ik}\}$ és el conjunt de valors de la variable
   i $Jaccard(X_{ik}, X_{jk})$ és el coeficient de semblança de Jaccard

   quan la variable $X_{.k}$ és qualitativa no ordenada i emmagatzema un conjunt de p valors o modalitats qualitatives d'un màxim de q modalitats diferents ($1 \leq p \leq q$)

5. $D_{local}(X_{ik}, X_{jk}) = \frac{lev_{X_{ik}, X_{jk}}(length(X_{ik}), length(X_{jk})) - |length(X_{ik}) - length(X_{jk})|}{max(length(X_{ik}), length(X_{jk})) - |length(X_{ik}) - length(X_{jk})|}$

   on

$$lev_{a,b}(i,j) = \begin{cases} \max(i,j) & si\ \min(i,j) = 0 \\ \min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + \begin{cases} 1 & si\ a_i \neq b_j \\ 0 & si\ a_i = b_j \end{cases} \end{cases} & altrament \end{cases}$$

és la distància de Levenshtein entre els primers *i* caràcters de l'string *a* i els primers *j* caràcters de l'string *b*.

quan la variable $X_{.k}$ és un *string* en format lliure, que emmagatzema un sol valor que és un *string* qualsevol

Centroide d'un clúster

Donat un clúster $C_p = \left\{X^p_{1.}, X^p_{2.}, \dots, X^p_{N_p.}\right\}\ on\ N_p = \#C_p$, es defineix el centroide d'un clúster com el següent vector:

$Centroide\ (C_p) = [\overline{X^p_{.1}}, \overline{X^p_{.2}}, \dots, \overline{X^p_{.N}}\ ]$ on cada component es calcula de forma diferent depenent de la tipologia de la variable:

$$\overline{X^p_{.j}} = \begin{cases} \sum_{i=1}^{N_p} \frac{X^p_{ij}}{N_p} & si\ X^p_{.j} és\ una\ variable\ quantitativa\ o\ numèrica \\ moda(X^p_{.j}) & si\ X^p_{.j} és\ una\ variable\ qualitativa\ amb\ 1\ valor\ possible \\ \arg\max_{i \in 1..N_p} freq(X^p_{ij}) & si\ X^p_{.j} és\ una\ variable\ qualitativa\ amb\ més\ d'un\ valor\ possible \\ \arg\max_{i \in 1..N_p, l \in 1,..\#SemW^p_{ij}} freq\left\{SemW^p_{ij}(l)\right\} & si\ X^p_{.j} és\ una\ variable\ string\ en\ format\ lliure \end{cases}$$

On $\left\{SemW^p_{ij}(l)\right\} = Semantic\ Words\ \{X^p_{ij}\}_{l=1,..,\#SemW^p_{ij}}$ és el conjunt de paraules amb significat semàntic (noms) que hi ha a l'string.

Mesures d'avaluació de la qualitat d'un clustering

En clustering, com no disposem de mesures ni etiquetes (grups) de referència respecte les quals fer una avaluació (en ser aprenentatge no supervisat), cal utilitzar mètriques internes per estimar la qualitat dels resultats obtinguts. Aquestes mètriques es basen exclusivament en els nivell de cohesió i separació dels clusters.

Com a mínim us demanem que implementeu la mètrica de Coeficient de Silhouete, que mesura, per a cada individu, com n'és de similar als members del seu clúster en comparació amb els altres clústers:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

On $a(i)$ és la distància mitjana entre $a(i)$ i la resta d'elements del seu propi clúster (cohesió), i $b(i)$ és la distància mitjana mínima entre i i els elements de qualsevol altre clúster (separació).

El valor global del clustering es calcula com la mitjana:

$$S = \frac{1}{n}\sum_{i=1}^{n} s(i)$$

Si S s'aproxima a 1, es considera que els clústers ben separats i compactes; si s'aproxima a 0, s'interpreta que els individus estan massa propers a les fronteres entre clústers, i $S < 0$ indica que hi ha molts individus incorrectament assignats.

Com a alternativa o com a complement, podeu investigar i explorar els següents mètodes:

- Índex Calinski-Harabasz: es basa en avaluar la relació entre la dispersió inter-cluster i la dispersió intra-cluster.
- Índex Davies-Bouldin: quantifica la similitud entre clústers.

Determinació del nombre de clústers k

Com a funcionalitat opcional, podeu implementar un procediment de selecció automàtica de la k. Una opció és mitjançant el mètode del colze (*Elbow Method*), en el qual es calcula la variança explicada pels clústers obtinguts en funció de k i es tria la k on la reducció marginal s'estabilitza (el punt de "colze", veure figura).

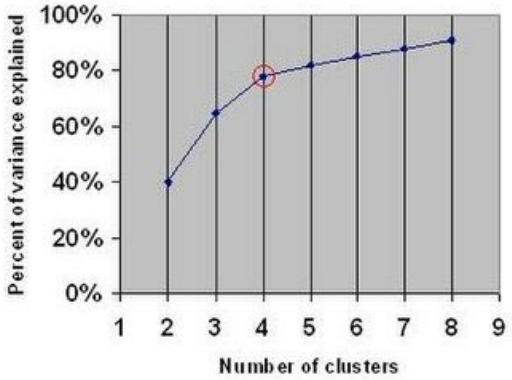

Figure 1. Punt de colze identificat (imatge de: https://es.wikipedia.org/wiki/Método_del_codo_(agrupamiento))